On Sample Optimality in Personalized Collaborative and Federated Learning

Mathieu Even[1], Laurent Massoulié [1,2], Kevin Scaman [1]

[1]Inria Paris - Département d'informatique de l'ENS, PSL Research University

[2]Microsoft-Inria Joint Center

Abstract

In personalized federated learning, each member of a potentially large set of agents aims to train a model minimizing its loss function averaged over its local data distribution. We study this problem under the lens of stochastic optimization, focusing on a scenario with a large number of agents, that each possess very few data samples from their local data distribution. Specifically, we prove novel matching lower and upper bounds on the number of samples required from all agents to approximately minimize the generalization error of a fixed agent. We provide strategies matching these lower bounds, based on a *gradient filtering* approach: given prior knowledge on some notion of distance between local data distributions, agents filter and aggregate stochastic gradients received from other agents, in order to achieve an optimal bias-variance trade-off. Finally, we quantify the impact of using rough estimations of the distances between local distributions of agents, based on a very small number of local samples.

1 Introduction

A central task in federated learning [30, 39] is the training of a common model from local data sets held by individual agents. A typical application is when users (*e.g.* mobile phones, hospitals) want to make predictions (*e.g.* next-word prediction, treatment prescriptions), but each has access to very few data samples, hence the need for collaboration. As highlighted by many recent works (*e.g.* Hanzely et al. [27], Mansour et al. [38]), while training a global model yields better statistical efficiency on the combined datasets of all agents by increasing the number of samples linearly in the number of agents, this approach can suffer from a dramatically poor generalization error on local datasets. A solution to this generalization issue is the training of *personalized* models, a midway between a shared model between agents and models trained locally without any coordination.

An ideal approach would take the best of both worlds: increased statistical efficiency by using more samples, while keeping local generalization errors low. This raises the fundamental question: what is the optimal bias/variance tradeoff between personalization and coordination, and how can it be achieved?

We formulate the personalized federated learning problem as follows, studying it under the lens of stochastic optimization [5]. Consider $N \in \mathbb{N}^*$ agents denoted by integers $1 \leqslant i \leqslant N$, each desiring to minimize its own local function $f_i : \mathbb{R}^d \to \mathbb{R}$, while sharing their stochastic gradients. Since only a limited number of samples are locally available, we focus on *stochastic gradient descent*-like algorithms, where agents each sequentially compute stochastic gradients g_i^k such that $\mathbb{E}\left[g_i^k\right] = \nabla f_i$. In order to reduce the sample complexity, *i.e.* the number of samples or stochastic gradients required to reach small generalization error, agents thus need to use stochastic gradients from other agents, that are *biased* since in general $\mathbb{E}\left[g_i^k\right] \neq \nabla f_j$. Our algorithms are based on a *gradient filtering* approach:

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

upon reception of stochastic gradients $(g_j^k)_j$, agent i *filters* these gradients and aggregates them using some weights λ_j into $\sum_j \lambda_j g_j^k$, in order to achieve some bias/variance trade-off.

1.1 Contributions and outline of the paper

In this paper, we consider an oracle model where at each step $k = 1, 2, \ldots$, all agents may draw a sample according to their local distribution. We aim at computing the number of stochastic gradients sampled from all agents, required to reach a small generalization error, in terms of biases (distances between functions or distributions), regularity, and noise assumptions. The oracle model, main assumptions and problem formulations are given in Section 2. Our main contributions are then as follows.

(i) In Section 3 we prove *information theoretic* lower bounds: to reach a target generalization error $\varepsilon > 0$ for a fixed agent i, no algorithm can achieve a reduction in the number of oracle calls by a factor larger than the total number of agents ε-close –in a suitable sense– to agent i.

(ii) We next study a naive strategy based on weighted gradient averaging algorithms, coined *all-for-one*, that matches this lower bound, at the cost of high communication and storage requirements.

(iii) We then propose in Section 5 a parallel extension of the simple weighted gradient averaging algorithm that yields an efficient algorithm for collaborative generalization error minimization problems. In this algorithm, agents compute stochastic gradients at *their* local estimate, and broadcast it to other agents who may use these to update their own estimates. For $x^k = (x_1^k, \ldots, x_N^k)$ where x_i^k is the local estimate of agent i at iteration k, updates of the ALL-FOR-ALL algorithm write as:

$$x^{k+1} = x^k - \eta W g^k \,,$$

where $g^k = (g_1^k, ..., g_N^k)$ for an unbiased stochastic gradient g_i^k of function f_i, a step size η, and a carefully chosen symmetric matrix W. Agent i thus uses stochastic gradients that are doubly biased, as gradients of a "wrong function" f_j instead of f_i computed at a "wrong location" x_j^k instead of x_i^k. Interestingly, note that the ALL-FOR-ALL algorithm is not a gossip algorithm *per se* (see *e.g.* [44]), since the matrix W is not doubly-stochastic: gradients are not aggregated with weights that sum to 1. Moreover, W depends on the distance between local agents distributions, and thus requires either prior information on the local distributions, or estimating these distances as a pre-processing step.

(iv) We finally study in Section 6 the impact of estimating, based on a very limited number of samples, the matrix W to use in the ALL-FOR-ALL algorithm. Under a mixture model assumption on the agents, we obtain that for a bounded – up to logarithmic factors – number of samples per agent, any arbitrary small generalization can be reached, with an optimal collaboration speedup in terms of the number of agents in each mixture of the mixture model.

1.2 Related works

Federated Learning is a paradigm in machine learning where training is done collaboratively among several agents, taking into account privacy constraints [30, 35, 39, 50]. A central task is the training of a common model for all agents, for which both *centralized* approaches orchestrated by a server and *decentralized* approaches with no central coordinator [43] have been considered. The algorithms we propose in this paper are well suited for a decentralized implementation.

As observed in Hanzely et al. [27], training a common model for all users can lead to poor generalization on certain tasks such as e.g. next-word prediction. To improve both accuracy and fairness, *personalized* models thus need to be learnt for each agent [37, 42, 53]. Approaches to personalization include fine-tuning [10, 33], transfer learning techniques [18, 48, 50], using shared-representation models [11]. Personalization in FL can also be formulated as the training of local models with a regularization term that enforces collaboration between users [27] or with a meta-learning approach [9, 24, 29]. We refer the interested reader to Kulkarni et al. [36] for a broader survey of Personalized Federated Learning.

While the goal of personalization is to minimize local generalization errors, the above cited works do not provide theoretical guarantees over the sample complexity to obtain small local errors, but instead control errors on a regularized problem, in terms of communication rounds or full gradients used, and not in terms of samples used. Deng et al. [14], Mansour et al. [38] among others provide generalization errors under a statistical learning framework that depend on VC-dimensions and

on distances between each local data distribution and the mixture of all datasets. Donahue and Kleinberg [15, 16] study the bias-variance trade-off between collaboration and personalization for mean estimation in a game-theoretic framework. Beaussart et al. [3], Chayti et al. [8], Grimberg et al. [26] also concurrently frame personalization as a stochastic optimization problem with biased gradients and are the works closest to ours. They consider the training of a single agent with biased gradients from another group of agents dedicated to this agent, and obtain performance guarantees in terms of distance between individual function f_i and the average $N^{-1}\sum_j f_j$. In contrast, we obtain more general performance bounds based on distance bounds between all pairs of functions f_i, f_j (or equivalently, pairs of local distributions), in the case where all agents desire to minimize their local objective; our "bias assumption" is also milder. In addition, we prove matching lower bounds, and study under a mixture model the statistical efficiency of our approach.

Finally, data-heterogeneity has long been a challenge in Federated optimization, as for instance noted in the analyses and performances of the Local SGD algorithm [32, 52]. Many algorithmic solutions have been proposed to counterweight this effect [31, 41] (non-exhaustive list). Yet this line of work studies the effect of data-heterogeneity on the convergence guarantees of FL algorithms that train *one* global model, irrespectively of the local generalization property of this trained global model. Our work is orthogonal, and focuses on data-heterogeneity as a challenge for statistical meaning (local generalization) of the model(s) trained, as opposed to related works that study data-heterogeneity as a challenge in distributed or federated learning to design fast and scalable algorithms. Putting into perspective these two views on the challenge data-heterogeneity in FL seems however necessary, and stresses its importance.

2 Problem Statement and Assumptions

We now detail our objectives and the necessary technical assumptions. We consider general stochastic gradient methods and formulate our problem, assumptions and algorithms accordingly.

2.1 Problem setting

Let $\mathcal{D}_i$ for $1 \leqslant i \leqslant N$ be a probability distribution on a set Ξ (agent i's local distribution, *not* its empirical distribution), $\ell : \mathbb{R}^d \times \Xi \to \mathbb{R}$ a loss function. We assume that the function f_i that agent i aims at minimizing is the generalization error on agent i's local distribution:

$$f_i(x) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[\ell(x, \xi_i)\right] , \quad x \in \mathbb{R}^d . \tag{1}$$

We coin this problem as *collaborative generalization error minimization (GEM)*. At every iteration $k = 1, 2, \ldots$, agent i may access unbiased *i.i.d.* estimates $g_i^k(x)$ of $\nabla f_i(x)$:

$$g_i^k(x) = \nabla_x \ell(x, \xi_i^k) , \quad \xi_i^k \sim \mathcal{D}_i , \quad x \in \mathbb{R}^d , \, 1 \leqslant i \leqslant N .$$

Counting the number of stochastic gradients used in the whole set of agents to reach a precision ε for f_i thus reduces to computing the number of samples required from all agents to obtain local generalization error ε for agent i. To specify the information shared between agents via access to stochastic gradients, we define the following oracle, that lets at every iteration all agents sample a stochastic gradient. After K oracle queries, each agent will have sampled K stochastic gradients for a total of NK in the whole set of agents. Let $\left\{(\xi_1^k, \ldots, \xi_N^k), k \geqslant 0\right\}$ a sequence of *i.i.d.* random variables of law $\mathcal{D}_1 \times \cdots \times \mathcal{D}_N$. Given the initial shared knowledge $\mathcal{S}_0$, at iterations $k = 1, 2, \ldots$,

1. For all $1 \leqslant j \leqslant N$, agent j samples ξ_j^k chooses some $y_j^k \in \mathbb{R}^d$ as a $\mathcal{S}_{k-1}$-measurable function.

2. The shared memory is extended: $\mathcal{S}_k = \mathcal{S}_{k-1} \cup \left\{g_j^k(y_j^k), \xi_j^k \, , \, 1 \leqslant j \leqslant N\right\}$.

3. Agent j outputs x_j^k as a $\mathcal{S}_k$-measurable function.

For fixed target precision $\varepsilon > 0$, the objective is to find, using T_ε samples from all agents in total - corresponding to $K_\varepsilon = T_\varepsilon / N$ oracle calls -, models with local generalization error ε. Throughout the paper, we assume that each function f_i is minimized over $\mathbb{R}^d$, and we denote by $x_i^\star$ such a minimizer. We further consider the following two standard assumptions.

Assumption 1 (Noise)**.** *There exists $\sigma^2 > 0$ such that for all $1 \leqslant i \leqslant N$ and $x \in \mathbb{R}^d$,*

$$\mathbb{E}_{\xi_i \sim \mathcal{D}_i} \|\nabla_x \ell(x, \xi_i) - \nabla f_i(x)\|^2 \leqslant \sigma^2 .$$

Assumption 2 (Regularity)**.** *Functions f_i are μ-strongly convex and L-smooth [7].*

2.2 Distribution-based distances

We first introduce extensions of classical *Integral Probability Metrics* (IPMs, [46]) to multivariate functions, i.e. pseudo-distances on the set of probability measures parameterized by a set $\mathcal{H}$ of functions, fixed in the sequel.

Definition 1. *For $\mathcal{H}$ a set of functions from Ξ to $\mathbb{R}^d$ and $\mathcal{D}, \mathcal{D}'$ two probability distributions on Ξ, we define:*

$$d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') = \sup_{h \in \mathcal{H}} \left\| \mathbb{E}\left[h(\xi) - h(\xi')\right] \right\|,$$

where $\xi \sim \mathcal{D}$ and $\xi' \sim \mathcal{D}'$. $d_{\mathcal{H}}$ is a pseudo-distance on the set of probability measures on Ξ.

This family of pseudo-distances contains a large number of standard distances between distributions, including total variation (with the set of 1-locally bounded functions, functions that send any ball of radius 1 in a ball of radius 1), the Wassertein distance (with the set of 1-Lipschitz functions), maximum mean discrepancies (with the unit ball of a RKHS), or even a simple distance between means of the distributions (with the set of 1-Lipschitz affine functions), developed further in Section 6.

Assumption 3 (Distribution-based dissimilarities)**.** *For some non-negative weights $(b_{ij})_{1 \leqslant i,j \leqslant N}$, we have $d_{\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_j) \leqslant b_{ij}$ for all $1 \leqslant i, j \leqslant N$. We further assume that either of the following holds.*

1. (Weak dissimilarities). *For all $1 \leqslant i \leqslant N$, $\big(\xi \in \Xi \mapsto \nabla_x \ell(x_i^\star, \xi)\big) \in \mathcal{H}$.*
2. (Strong dissimilarities). *For all $x \in \mathbb{R}^d$, $\big(\xi \in \Xi \mapsto \nabla_x \ell(x, \xi)\big) \in \mathcal{H}$.*

The "weak dissimilarities" assumption is of course easier to satisfy than the "strong" version, and our results will ultimately depend only on the weak assumption. Under Assumption 3 (weak version) and Assumption 2, we have $f_i(x_j^\star) - f_i(x_i^\star) \leqslant b_{ij}^2/(2\mu)$, which motivates our use of distribution-based dissimilarity assumptions.

Notation: in the rest of the paper, variables t or T denote the number of stochastic gradients g_i^k sampled (or data item sampled from personal distribution) from all agents, while variables k or K denote the iterates of the algorithms (or equivalently the number of oracle calls made).

3 Information-theoretic lower bound on the sample complexity

In this section, we prove lower bounds on the total number of stochastic gradients required from all agents, to reach ε-generalization for a given agent. Our lower bounds apply to collaborative GEM, i.e. functions $(f_i)_{1 \leqslant i \leqslant N}$ of the form (1), for shared loss function ℓ and user distributions $\mathcal{D}_1, \ldots, \mathcal{D}_N$.

An oracle $\phi : \mathbb{R}^{N \times d} \to \mathcal{I}$ is a random function that answers some $\phi(x) \in \mathcal{I}$ where $\mathcal{I}$ is an information set, for every query $x \in \mathbb{R}^{N \times d}$. We adapt the definitions of Agarwal et al. [1] of sample complexity for *SGD* to our personalization problem. Formally, the *first-order* oracle we defined in Section 2 and that we write as $\phi\big((\mathcal{D}_i)_{i=1,\ldots,N}, \ell\big)$ for shared loss function ℓ and user distributions $\mathcal{D}_1, \ldots, \mathcal{D}_N$, returns for $x \in \mathbb{R}^{N \times d}$:

$$\phi\big((\mathcal{D}_i)_i, \ell\big)(x) = \Big(i,\, x_i,\, \xi_i,\, \ell(x_i, \xi_i),\, g_i^k(x_i)\Big)_{1 \leqslant i \leqslant N},$$

where $\xi_i \sim \mathcal{D}_i$. Given distributions and a loss function $\big((\mathcal{D}_i)_i, \ell\big)$, we denote by $\mathbb{M}$ the set of all methods $\mathcal{M} = (\mathcal{M}_K)_{K \geqslant 0}$: for any $K \geqslant 0$, $\mathcal{M}_K$ makes K oracle calls from oracle $\phi\big((\mathcal{D}_i)_i, \ell\big)$ (while using $T = NK$ stochastic gradient samples from all agents), and returns $x_i^K \in \mathbb{R}^d$ for agent i as a measurable function of the K oracle calls. For a set $\mathbb{D}$ of couples of distributions and loss function $\big((\mathcal{D}_j)_j, \ell\big)$ defining functions $(f_i)_{1 \leqslant i \leqslant N}$, we are interested in lower-bounding:

$$\inf_{\mathcal{M} \in \mathbb{M}} \sup_{((\mathcal{D}_j)_j, \ell) \in \mathbb{D}} \mathcal{K}_i^\varepsilon\Big(\mathcal{M}, \big((\mathcal{D}_j)_j, \ell\big)\Big),$$

where $\mathcal{K}_i^\varepsilon\Big(\mathcal{M}, \big((\mathcal{D}_j)_j, \ell\big)\Big)$ is the number of oracle calls required to reach generalization error $\varepsilon > 0$ for agent i, and writes as:

$$\mathcal{K}_i^\varepsilon\Big(\mathcal{M}, \big((\mathcal{D}_j)_j, \ell\big)\Big) = \inf\left\{K \in \mathbb{N}^* \text{ such that } \mathbb{E}\left[f_i(x_i^K) - \min_{x \in \mathbb{R}^d} f_i(x)\right] \leqslant \varepsilon\right\}.$$

We now define the set $\mathbb{D}$ we consider for our lower bounds. Let $b = (b_{ij})_{1 \leqslant i,j \leqslant N}$ be non-negative weights that verify the triangle inequality – namely, $b_{ij} \leqslant b_{ik} + b_{kj}$ for all i, j, k –, and let $r, \mu, L, \sigma > 0$. $\mathbb{D}_\mu^L(r, b, \sigma)$ is the set of all $\big((\mathcal{D}_i)_{1 \leqslant i \leqslant N}, \ell\big)$, such that the functions f_i parameterized by these tuples of distributions and shared loss function verify Assumptions 1, 2 and 3 for σ^2, $\mu, L > 0$ and b, such that $\|x_i^\star\| \leqslant r$ for all $1 \leqslant i \leqslant N$, and such that $f_i(x_j^\star) - f_i(x_i^\star) \leqslant b_{ij}^2/(2\mu)$. We use the notation $a(\cdot) = \Omega(b(\cdot))$ for $\exists C > 0$ such that $a(\cdot) \geqslant C b(\cdot)$.

Theorem 1 (IT lower bound)**.** *Let $\varepsilon \in (0, 1/16)$, (b_{ij}) verifying the triangle inequality, $r, \sigma > 0$. Assume that the function set $\mathcal{H}$ contains the all 1-Lipschitz affine functions and that $d_{\mathcal{H}} \leqslant d_{\mathrm{TV}}$. For some constant $C > 0$ independent of the problem and any $i \in \{1, \dots, N\}$:*

$$\inf_{\mathcal{M} \in \mathbb{M}} \sup_{((\mathcal{D}_j)_j, \ell) \in \mathbb{D}_{\mu=1/r^2}^{L=1/r^2}(r,b,\sigma)} \mathcal{K}_i^\varepsilon\Big(\mathcal{M}, \big((\mathcal{D}_j)_j, \ell\big)\Big) = \Omega\left(\frac{r^2\sigma^2}{\varepsilon \mathcal{N}_i^\varepsilon(\frac{b^2}{4\mu})}\right),$$

where $\mathcal{N}_i^\varepsilon(\frac{b^2}{4\mu}) = \sum_j \mathbb{1}_{\{b_{ij}^2 \leqslant 4\mu\varepsilon\}}$ is the number of agents j verifying $b_{ij}^2 \leqslant \varepsilon$.

The proof of this lower bound (Appendix B) builds on lower bounds based on Fano's inequality [20] for stochastic gradient descent [1] or for information limited statistical estimation [19, 54], adapted to personalization. Theorem 1 states that, given the knowledge of (b_{ij}), σ^2, $\mu = 1$ and $L = 1$, there exist difficult instances of the problem that satisfy all three Assumptions 1, 2 and 3, such that the number of oracle calls needed to obtain a generalization error of ε for an agent i is lower-bounded by the right hand side of the equation in Theorem 1.

The factor $C\sigma^2 r^2 \varepsilon^{-1}$ is reminiscent of stochastic gradient descent, and is present in Agarwal et al. [1]: without cooperation, this is the sample complexity of *SGD* for a fixed agent. Cooperation appears in the factor $1/\mathcal{N}_i^\varepsilon(b/4)$: the sample complexity is inversely proportional to the number of agents j that have distributions similar to that of i. One cannot hope for better than a linear *collaboration speedup* proportional to agents $4\mu\varepsilon$-close to i in terms of the distance $d_{\mathcal{H}}$. Theorem 1 is a *worst-case* lower bound, so that a collaboration speedup could be leveraged even for small ε, but this would require making stronger additional assumptions.

4 The ALL-FOR-ONE algorithm: parallel weighted gradient averagings

After providing lower complexity bounds in Theorem 1, we present in this section a naive algorithmic approach based on weighted gradient averagings (WGA), that proves to be sample-optimal. Each agent i keeps N shared local models $x_1^k, \dots, x_N^k$, where x_j^k estimates $x_j^\star$ at iteration k (the knowledge of x_j^k needs to be shared by all agents). At each iteration k, when a sample ξ_j^k is obtained at agent j, it is used by that agent to compute unbiased estimates of $\nabla f_j(x_i^k)$ for all $i \in [N]$. The iterates of the WGA algorithm write as, where $\lambda_{ij} \geqslant 0$ are such that $\sum_j \lambda_{ij} = 1$ for all $1 \leqslant i \leqslant N$:

$$x_i^{k+1} = x_i^k - \eta \sum_{j=1}^N \lambda_{ij} \nabla f_j(x_i^k, \xi_j^k), \tag{2}$$

for some step size $\eta > 0$. We call this algorithm that consists in performing N parallel WGA algorithms ALL-FOR-ONE (AFO), since every iteration of each gradient averaging for a given node i requires all the other nodes to compute one stochastic gradient for i. WGA is thus equivalent to training models on the mixture of distributions $(\mathcal{D}_j)_j$ with weights $(\lambda_{ij})_j$ for all i.

Theorem 2. *Let $(x_i^k)_{1 \leqslant i \leqslant N,\, k \geqslant 0}$ be generated with (2), and assume that Assumptions 1, 2 and 3 (strong version) hold. For any $K \geqslant 0$ and $1 \leqslant i \leqslant N$, and for η as in Equation (6),*[1]

$$\mathbb{E}\left[f_i(x_i^K) - f_i(x_i^\star)\right] \leqslant \big(f_i(x_i^{(0)}) - f_i(x_i^\star)\big) e^{-\frac{K}{2\kappa}} + \tilde{\mathcal{O}}\left(\frac{\kappa\sigma^2}{\mu K} \sum_{1 \leqslant j \leqslant N} \lambda_{ij}^2\right) + \sum_{1 \leqslant j \leqslant N} \lambda_{ij} \frac{b_{ij}^2}{\mu}.$$

[1] $\tilde{\mathcal{O}}$ hides logarithmic and constant factors

Let $\varepsilon > 0$. For a specific choice of $\lambda_{ij} = \frac{\mathbb{1}_{\{b_{ij}^2 < \varepsilon/2\}}}{\mathcal{N}_i^\varepsilon(2b)}$, WGA (2) satisfies $\mathbb{E}\left[f_i(x_i^{K_\varepsilon(i)}) - f_i(x_i^\star)\right] \leqslant \varepsilon$ for a number of oracle calls of:

$$K_i(\varepsilon) = \tilde{\mathcal{O}}\left(\frac{\kappa\sigma^2}{\mu\varepsilon}\frac{1}{\mathcal{N}_i^\varepsilon(2b^2/\mu)}\right),$$

where $\mathcal{N}_i^\varepsilon$ is previously defined in Theorem 1.

Since the oracle complexity of the WGA algorithm matches that of our lower bound, this proves that our lower bound is optimal. However, this algorithm may be difficult to use in practice: *(i)* the choice of λ_{ij} is an explicit function of distribution distances (b_{ij}) (defined in Assumption 3) that can be (statistically speaking) as hard to compute as solving our optimization problem; and *(ii)* the memory requirements and computation/communication costs of WGA can be prohibitive for large N and large ε (they scale with $\mathcal{N}_i^\varepsilon$ for agent i). Note that the strong version of Assumption 3 used in Theorem 2 can be replaced by a more classical uniform bound of the form $\|\nabla f_i - \nabla f_j\| \leqslant b_{ij}$.

We first begin by solving this latter issue – an algorithmic one – in the next section, by introducing and studying the ALL-FOR-ALL algorithm. We discuss *(i)* in Section 6, where we provide scenarii over which statistical theoretical guarantees can be derived on the error made by estimating these distribution distances using only a few samples.

5 The ALL-FOR-ALL algorithm

Algorithm 1 *All-for-all* algorithm

1: Step size $\eta > 0$, matrix $W \in \mathbb{R}^{N\times N}$, initialization $x_1^0 = \ldots = x_N^0 \in \mathbb{R}^d$ (x_i^0 at agent i).
2: **for** $k = 0, 1, 2, \ldots K-1$ **do**
3: Agents $1 \leqslant j \leqslant N$ compute $g_j^k(x_j^k)$ and broadcast it to all agents i such that $W_{ij} > 0$.
4: For $i = 1, \ldots, N$, update:

$$x_i^{k+1} = x_i^k - \eta \sum_{j\,:\,W_{ij}>0} W_{ij} g_j^k(x_j^k)\,.$$

5: **end for** Return x_i^K for agent i

In this section, we present the ALL-FOR-ALL algorithm (AFA), an adaptation of the weighted gradient averaging algorithm. For $1 \leqslant i \leqslant N$, initialize $x_i^0 = x_0 \in \mathbb{R}^d$. At iteration k, let $x_i^k \in \mathbb{R}^d$ be agent i's current estimate of $x_i^\star$, and denote $x^k = (x_i^k)_{1\leqslant i\leqslant N} \in \mathbb{R}^{N\times d}$. For a step size $\eta > 0$ and a matrix $W \in \mathbb{R}^{N\times N}$ with non-negative entries (remarkably and as discussed later, W will not necessarily verify $\sum_j W_{ij} = 1$), iterates of the *all-for-all* algorithm are generated with Algorithm 1. In Theorem 3, we control the averaged local generalization error amongst all agents:

$$F^k = \frac{1}{N}\sum_{i=1}^N f_i(x_i^k) - f_i(x_i^\star)\,, \quad k \geqslant 0\,.$$

Theorem 3 (ALL-FOR-ALL algorithm)**.** *Let $K > 0$, $\eta > 0$, and W a matrix of the form $W = \Lambda\Lambda^\top$ for some stochastic matrix $\Lambda = (\lambda_{ij})_{1\leqslant i,j\leqslant N}$. Assume that Assumptions 1, 2 and 3 (weak version) hold. The iterates $(x_i^k)_{k\geqslant 0, 1\leqslant i\leqslant N}$ generated with Algorithm 1 verify, for η as in Equation (7):*

$$\mathbb{E}\left[F^K\right] \leqslant F^0 e^{-\frac{K}{2\kappa}} + \tilde{\mathcal{O}}\left(\frac{\kappa\sigma^2}{K\mu N}\sum_{1\leqslant i,j\leqslant N}\lambda_{ij}^2\right) + \frac{1}{N}\sum_{1\leqslant i,j\leqslant N}\lambda_{ij}\frac{b_{ij}^2}{2\mu}\,.$$

As for the AFO algorithm, we can deduce from this result the number of oracle calls required by the ALL-FOR-ALL algorithm to reach an averaged ε-generalization, under the idealistic setting where the distribution-based distances b_{ij} are accessible.

Corollary 1. *Let $\varepsilon > 0$. Under the same assumptions as in Theorem 3, for a choice of matrix $W = \Lambda\Lambda^\top$ where $\lambda_{ij} = \frac{\mathbb{1}_{\{b_{ij}^2/\mu < \varepsilon\}}}{\mathcal{N}_i^\varepsilon(b^2/\mu)}$, the* ALL-FOR-ALL *algorithm (Algorithm 1) returns $(x_i^K)_{1 \leqslant i \leqslant N}$ satisfying $\frac{1}{N}\sum_{i=1}^N f_i(x_i^{K_\varepsilon}) - f_i(x_i^\star) \leqslant \varepsilon$, for a number of oracle calls satisfying:*

$$K_\varepsilon \leqslant 2 \max\left(\frac{\kappa\sigma^2}{\varepsilon\mu}\frac{1}{N}\sum_{i=1}^N \frac{1}{\mathcal{N}_i^\varepsilon(b^2/\mu)}, \kappa\right) \ln\left(\varepsilon^{-1}F^0\right).$$

Denoting $K_\varepsilon(i)$ the oracle complexity of the WGA algorithm - that matches the lower bound -, we observe that the ALL-FOR-ALL algorithm reaches an averaged ε-generalization with an number of oracle calls K_ε, of $K_\varepsilon \leqslant \frac{1}{N}\sum_i K_\varepsilon(i)$. The speedup in comparison with a no-collaboration strategy (all agents locally performing SGD) is $\frac{1}{N}\sum_i \frac{1}{\mathcal{N}_i^\varepsilon(b^2/\mu)}$: the mean of all local speedups.

Remark 1. *In Theorem 3, as its proof shows, the quantities $b_{ij}^2/(2\mu)$ in the last term can in fact be replaced by the quantities $f_i(x_j^\star) - f_i(x_i^\star)$, that control how well the optimal model for j generalizes for i, and the bias induced by* ALL-FOR-ALL *iterations is a weighted average of these quantities. Note that in our lower bound (Theorem 1), we enforce that the functions considered are required to satisfy $f_i(x_j^\star) - f_i(x_i^\star) \leqslant b_{ij}^2/(2\mu)$. We believe this notion of function proximity that we leverage to be the weakest achievable in our setting; no prior work uses such a mild proximity assumption.*

Perhaps surprisingly, matrix W is in general *not* a gossip matrix (*i.e.* such that $W\mathbb{1} = \mathbb{1}$): agent i does not aggregate a convex combination of stochastic gradients, but a combination with scalars that do not necessarily sum to 1. We thus cannot say that the ALL-FOR-ALL algorithm acts as if, in parallel, each agent i trains a model on the mixture of distributions $\mathcal{D}_j$ with weights W_{ij}. In fact, as the analysis shows below, agent i trains a model on the mixture of distributions, with weights λ_{ij}, if Λ is a stochastic square root of matrix W ($\Lambda\Lambda^\top = W$), as in the AFO algorithm. In order to account for inter-dependencies between agents that do not directly share information, the *all-for-all gradient filtering* uses weights W_{ij} to aggregate information, instead of λ_{ij}. Propagating information using a matrix W, that induces a similarity graph G_W on $\{1, \ldots, N\}$, such that $(ij) \in E_W$ if $W_{ij} > 0$, is quite natural [4, 49]; yet, ours is the first analysis to give such precise generalization error bounds, through the use of a stochastic optimization framework.

In comparison to Theorem 3, the classical personalized FL approaches that consider personalized local models of the form $x_i = \bar{x} - \delta_i$, where $\bar{x}$ is some global quantity shared by all agents, perturbed (and personalized) by some local quantity δ_i (*e.g.* averaging between local and a global models), can be seen as the special instances where, for all i, we have $\lambda_{ii} = 1 - \alpha_i$ and $\lambda_{ij} = \frac{\alpha_i}{N-1}$ if $i \neq j$ for some α_i, and leads to bias terms of the form $\frac{1}{N}\sum_i \frac{\alpha_i}{N-1}\sum_{j \neq i} b_{ij}$ [14, 24, 38]. Full and naive collaboration (a single model trained for all users) corresponds to $\lambda_{ij} = 1/N$ for all i, j, and leads to a bias term of $\frac{1}{N^2}\sum_{i,j} b_{ij}$. The degrees of freedom offered by our matrix W (and by coefficients λ_{ij}) enable pairwise agent adaptation, and tighter generalization guarantees and bias/variance tradeoffs.

Proof sketch of Theorem 3. Since brutally analyzing convergence of the iterates (x^k) generated with $x^{k+1} = x^k - WG^k$ seems impossible due to both gradient biases and model biases between agents, we study these iterates through the introduction of a different but related problem. This approach is in fact similar to some decentralized optimization ones, where a dual problem or a related energy function is often introduced [22, 44], upon which well-studied algorithms are applied. The related problem we formulate is different from and more flexible than all the different personalized FL problems in the literature [27, 47], that consider regularization terms that enforce consensus. For $\lambda = (\lambda_{ij})_{1 \leqslant i,j \leqslant N}$ a stochastic matrix (such that for all $1 \leqslant i \leqslant N$, we have $\sum_{j=1}^N \lambda_{ij} = 1$), let f^Λ be defined as:

$$f^\Lambda(y) = \bar{f}(\Lambda y), \quad y \in \mathbb{R}^{N \times d}, \tag{3}$$

where $\bar{f} = \frac{1}{N}\sum_i f_i$. Gradient descent on f^Λ writes as $y^{k+1} = y^k - \eta\Lambda^\top\nabla\bar{f}(\Lambda y^k)$, where $\nabla\bar{f}(x) = \frac{1}{N}\big(\nabla f_i(x_i)\big)_{1 \leqslant i \leqslant N}$ for any $x \in \mathbb{R}^{N \times d}$. Importantly, notice that denoting $x^k = \Lambda y^k$ and since $W = \Lambda\Lambda^\top$, we have the recursion $x^{k+1} = x^k - \eta W\nabla\bar{f}(x^k)$, making an analysis of the iterates (x^k) possible. In our case, we however use stochastic gradients given by our oracle. The full gradient $\nabla\bar{f}(x)$ is thus replaced by $(g_i^k(x_i))_i$. Defining $(y^k)_k$ with the recursion: $y^{k+1} = y_k - \eta((\Lambda G^k(y^k)_i)_i$, initialized at $y_1^0 = x_1^0 = \ldots = y_N^0 = x_N^0$ we have $x^k = \Lambda y^k$, for all $k \geqslant 0$, where (x^k) is generated

using Algorithm 1. As a consequence, controlling in Theorem 3 the function values $\frac{1}{N}\sum_i f_i(x_i^k)$ is equivalent to controlling $f^\Lambda(y^k)$ (these two quantities are equal). The bias-variance trade-off thus writes as, where y^Λ minimizes f^Λ and $x^\star = (x_i^\star)_{1\leqslant i\leqslant N}$:

$$F^k \leqslant \underbrace{f^\Lambda(y^\Lambda) - \bar{f}(x^\star)}_{\textit{Bias term}} + \underbrace{f^\Lambda(y^k) - f^\Lambda(y^\Lambda)}_{\textit{Optimization and variance terms}} .$$

The rest of the proof, deferred to Appendix D, consists in relating these two terms to Λ and b. □

After providing the optimization tools and results to answer for the shortcomings of weighted gradient averagings, we now turn to quantifying the impact of the use of estimated values $\hat{b}_{ij}$ of b_{ij}, in order to close the loop.

6 Estimation of $d_{\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_j)$ as a pre-processing step

The sample complexity of estimating the distances $d_{\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_j)$ depends on the complexity of the function space $\mathcal{H}$. While the estimation of Wassertein or total variation distances are usually hard (in $\mathcal{O}(1/S^{1/d})$ where d is the ambient dimension and S the number of samples available for the estimation, see *e.g.* [51]), maximum mean discrepancy (MMD) distances often exhibit lower sample complexities in $O(1/\sqrt{S})$ [46]. Moreover, explicit assumptions on the loss function can also provide low sample complexities, as shown below for quadratic loss functions. Yet, the results presented in this section can be generalized beyond linear models with squared losses, as long as concentration inequalities for controlling how far empirical distributions are from the true distribution in terms of distance $d_{\mathcal{H}}$.

In order to formulate statistical results for the estimation of the pairwise distribution-based distances $d_{\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_j)$, we need to make additional structural assumptions, on both $\mathcal{H}$ and the distributions. Inspired by Collins et al. [11], we focus on analyzing an instance of our general GEM setting for quadratic losses and linear models, under which the generalization error of a given agent i writes as:

$$f_i(x) = \frac{1}{2}\mathbb{E}\left[\left(a_i^\top x - b_i\right)^2\right] , \quad x \in \mathbb{R}^d ,$$

where $z_i = (a_i, b_i)$ is a random variable on $\mathbb{R}^d \times \mathbb{R}$. The stochastic gradients thus write as $\nabla_x \ell(x, \xi_i) = (a_i^\top x - b_i)a_i$ for $z_i = (a_i, b_i)$, and are thus linear functions of $\xi_i = z_i z_i^\top$. Hence, Assumption 3 (weak version) is satisfied for $\mathcal{H}$ the set of $D^\star$-Lipschitz and affine functions, where $D^\star$ bounds all $\|x_i^\star\|$ for $1 \leqslant i \leqslant N$, leading to:

$$d_{\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_j) \leqslant D^\star \|\mathbb{E}\left[\xi_i\right] - \mathbb{E}\left[\xi_j\right]\| .$$

We make the following assumption on the law $\mathcal{D}_i$ of the random variables ξ_i: they are non-isotropic subgaussian random variables, that thus benefit from concentration inequalities that are dimension-independent [21, 34].

Assumption 4. *For some non-negative symetric matrix Σ and all $1 \leqslant i \leqslant N$, ξ_i are centered and Σ-subgaussian:*

$$\mathbb{P}\left(\xi_i^\top y \geqslant u\right) \leqslant \exp\left(-\frac{u^2}{2y^\top \Sigma y}\right) , \quad \forall y \in \mathbb{R}^{(d+1)^2} , \quad \forall u > 0 ,$$

and we denote as ν^2 the largest eigenvalue of Σ, and $d_{\text{eff}} = \frac{\|\Sigma\|_2}{\nu^2}$ its effective dimension.

Importantly, note that d_{eff} can be arbitrarily smaller than the ambient dimension - for the MNIST dataset, d_{eff} is less than 3, while the ambient dimension is 712 [21]. Depending on a smaller dimension is also an assumption that Collins et al. [11] use in their work by exploiting shared representations.

We now formulate a structural assumption on the set of agents: there are M clusters $\mathcal{C}_1, \ldots, \mathcal{C}_M$ of C agents each (to ease notations, with a total number of agents $N = MC$). Within each cluster, agents distributions share the same objective, and clusters are "well-separated". These models are popular for modelling population heterogeneity and provide a formal framework for clustering problems; we refer the interested reader to Melnykov and Maitra [40] for a detailed survey on the subject.

Assumption 5 (Well-separated clusters of agents)**.** *For $M, C \geqslant 1$, N writes as $N = MC$ and there exists $\{\mathcal{C}_1, \dots, \mathcal{C}_M\}$ a partition of $\{1, \dots, N\}$, $\mu_1, \dots, \mu_m$ such that for all $1 \leqslant m \leqslant M$, $|\mathcal{C}_m| = C$ and for all $i, j \in \mathcal{C}_m$, we have $\mathbb{E}[\xi_i] = \mathbb{E}[\xi_j] = \mu_m$. We denote $\Delta^2 = \min_{m \neq m'} \|\mu_m - \mu_{m'}\|^2$ and assume that $\Delta^2 > 0$.*

When distribution-based distances were given (as in Corollary 1), Algorithm 1 achieved the optimal collaboration speedup, linear in $1/C$ under Assumption 5 and for small enough target precision ε. The cluster model is thus the natural baseline for our problem. In the case where agents estimate with whom to collaborate as we do in the sequel, reaching this collaboration speedup of $1/C$ will hence prove the effectiveness of the approach.

We assume that agents possess a limited number of samples. More precisely, for $1 \leqslant i \leqslant N$ and $S, K \geqslant 1$, agent i possesses $K + S$ *i.i.d.* samples of drawn from $\mathcal{D}_i$, S of which are dedicated to estimating who to collaborate with, the K remaining dedicated to the optimization process *i.e.* to running ALL-FOR-ALL iterations for a number K of oracle calls.

For $1 \leqslant i \leqslant N$, let $\hat{\mu}_i$ be an estimation of $\mathbb{E}[\xi_i]$ made with S *i.i.d.* samples $\xi_{i,1}, \dots, \xi_{i,S}$, and for $1 \leqslant i, j \leqslant N$ let $\hat{b}_{ij}$ be the following estimation of $d_{\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_j)$:

$$\hat{\mu}_i = \frac{1}{S} \sum_{s=1}^{S} \xi_{i,s}\,, \quad \hat{b}_{ij} = \|\hat{\mu}_i - \hat{\mu}_j\|\,.$$

Computing these distances can be done using only $\tilde{\mathcal{O}}(N)$ communications (rather than the N^2 communications of a naive approach) by performing randomized gossip communications [6] on the complete graph.

Theorem 4 (ALL-FOR-ALL with estimated biases)**.** *Assume that Assumptions 1, 2, 3 (for some unknown biases b_{ij}), 4 and 5 hold. Under the setting described, let $\hat{\Lambda}$ be the stochastic matrix with entries*

$$\hat{\lambda}_{ij} = \frac{\mathbb{1}_{\{\hat{b}_{ij}^2 \leqslant u\}}}{\sum_{\ell=1}^{N} \mathbb{1}_{\{\hat{b}_{i\ell}^2 \leqslant u\}}}\,,$$

for some $u > 0$ that verifies $u \geqslant \frac{4\nu^2 d_{\text{eff}}}{S}$. The ALL-FOR-ALL *algorithm with $W = \hat{\Lambda}\hat{\Lambda}^\top$ outputs $(x_i^K)_{1 \leqslant i \leqslant N}$ verifying, where $b_{\max} = \max_{i,j} b_{ij}$:*

$$\mathbb{E}\left[F^K\right] \leqslant F^0 e^{-\frac{K}{2\kappa}} + \tilde{\mathcal{O}}\left(\frac{\kappa\sigma^2}{K\mu}\left(\frac{1}{C} + C e^{-\frac{Su}{8\nu^2}}\right)\right) + \frac{2D^{\star 2} b_{\max} e^{-\frac{S\max(\Delta^2 - 2u, 2u^2)}{8\nu^2}} + 4u^2 \mathbb{1}_{\{2u \geqslant \Delta\}}}{\mu}\,,$$

where the mean is taken over both biases estimates $(\hat{b}_{ij})$ and gradient estimates (g_i^k).

Corollary 2. *Under the same assumptions as Theorem 4 and for $\varepsilon > 0$, the* ALL-FOR-ALL *algorithm with estimated biases as described above reaches an averaged generalization error of ε as long as:*

$$S = \tilde{\Omega}\Big(\frac{\nu^2 d_{\text{eff}}}{\Delta^2}\Big)\,, \quad C = \tilde{\Omega}\Big(\frac{\nu^2 d_{\text{eff}}}{\varepsilon}\Big)\,, \quad KC = \tilde{\Omega}\Big(\frac{\kappa\sigma^2}{\varepsilon\mu}\Big)\,, \quad K = \tilde{\Omega}(\kappa)\,.$$

Forgetting about the logarithmic factors, only a bounded number of local samples for each user (S and K) are required to reach an averaged arbitrarily small generalization error $\varepsilon > 0$, in the limit with an arbitrary large number of agents (N and C). Indeed, due to our regularity assumptions, K – the number of samples kept for the optimization problem – is required only to be of order κ, the condition number of the problem. The number of samples S used for estimating the biases is required to be of order $\nu_1^2 d_{\text{eff}}/\Delta^2$, the "signal-to-noise" ratio of our mixture model [40, 45], a natural quantity to depend on. Corollary 2 hence shows that the optimal collaboration speedup is achieved, up to logarithmic factors: in order to reach an arbitrary small generalization error $\varepsilon > 0$, are only required constant orders for S and K (the number of samples locally available) if the number of agents is large enough *i.e.* if $N = \tilde{\Omega}(M/\varepsilon)$, where M is the number of clusters *i.e.* we have a linear speedup in the clusters population. We numerically illustrate our theory in Appendix A on synthetic datasets, with clustered agents (as in this section), as well as in a setting where agents are distributed according to a more general "distribution of agent".

A closely related work [25] also studies a model where agents verify a cluster structure as described in Assumption 5 for quadratic losses and linear models. Yet, we highlight several differences between their approach and ours. First, Ghosh et al. [25] perform an *online* clustering of the agents, as opposed to our pre-training hierarchical approach. While the results we obtain in Theorem 4 and Corollary 2 and those of Ghosh et al. [25] have the same linear speedup in the number of agents, ours require no initialization condition. Finally, our algorithm is decentralized, thus leading to improved scalability (especially in terms of the number of clusters) and privacy [13], if of interest. Finally, not being restricted to clusters in the analysis of the ALL-FOR-ALL algorithm leads to a better collaboration speedup and fairness (in the sense that performance does not impact a few agents) in a non-clustered scenario, where an approach based on clusters would be highly non-optimal for agents that are at the border of the inferred clusters.

Conclusion

In this paper, we quantified in terms of function and distribution biases, stochastic gradient noise, target precision $\varepsilon > 0$ and functions regularity parameters, the benefit of collaboration between agents for shared minimization using stochastic gradient algorithms. Our lower bound (Theorem 1) states that, under prior knowledge on the distances between local distributions, the collaborative speedup can be linear only in the first phase of the optimization when the generalization error is large compared to the distances between distributions. More specifically, for a given agent i, the collaboration speedup is linear in the number of agents that are ε-close to i. Moreover, we show that the ALL-FOR-ONE algorithm allows such a speedup and is thus sample optimal. However, this algorithm requires high computation and communication capacities, a drawback that can be mitigated by the use of a novel algorithm called ALL-FOR-ALL, that benefits from the same collaboration speedup while being cheaper to deploy. Finally, we studied the impact of estimating distances between distributions as a pre-processing step to the optimization phase; under a mixture model assumptions on the agents, we obtain an optimal collaboration speedup. Extending our results – lower and upper complexity bounds – to other regularity assumptions and Section 6 to more general settings, as well as incorporating local steps in the ALL-FOR-ALL algorithm (even though to lighten communications, unbiased compressors could here be used, as our analysis encompasses these) are interesting questions left for future work. See [23] for extensions to convex-smooth functions (not necessarily strongly convex), convex-Lipschitz functions, and to asynchronous gradient oracles.

Aknowledgments This work was supported by ANR-19-P3IA-0001(PRAIRIE 3IA Institute). We also acknowledge support from the MSR-INRIA joint centre.

References

[1] Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

[2] Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

[3] Martin Beaussart, Felix Grimberg, Mary-Anne Hartley, and Martin Jaggi. WAFFLE: Weighted Averaging for Personalized Federated Learning. *arXiv:2110.06978 [cs]*, October 2021.

[4] Aurélien Bellet, Rachid Guerraoui, Mahsa Taziki, and Marc Tommasi. Personalized and private peer-to-peer machine learning. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 473–481, 2018.

[5] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, January 2018.

[6] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, June 2006. ISSN 0018-9448. doi: 10.1109/TIT.2006.874516. URL http://ieeexplore.ieee.org/document/1638541/.

[7] Sébastien Bubeck. Convex optimization: Algorithms and complexity, 2015.

[8] El Mahdi Chayti, Sai Praneeth Karimireddy, Sebastian U. Stich, Nicolas Flammarion, and Martin Jaggi. Linear speedup in personalized collaborative learning, 2021.

[9] Fei Chen, Mi Luo, Zhenhua Dong, Zhenguo Li, and Xiuqiang He. Federated meta-learning with fast convergence and efficient communication, 2018. URL https://arxiv.org/abs/1802.07876.

[10] Gary Cheng, Karan Chadha, and John Duchi. Fine-tuning is Fine in Federated Learning. *arXiv:2108.07313 [cs, math, stat]*, August 2021.

[11] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2089–2099. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/collins21a.html.

[12] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, April 2005.

[13] Edwige Cyffers, Mathieu Even, Aurélien Bellet, and Laurent Massoulié. Muffliato: Peer-to-peer privacy amplification for decentralized optimization and averaging, 2022. URL https://arxiv.org/abs/2206.05091.

[14] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive Personalized Federated Learning. *arXiv:2003.13461 [cs, stat]*, November 2020. arXiv: 2003.13461.

[15] Kate Donahue and Jon Kleinberg. Model-sharing games: Analyzing federated learning under voluntary participation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6): 5303–5311, May 2021.

[16] Kate Donahue and Jon Kleinberg. Optimality and stability in federated learning: A game-theoretic approach. In *Advances in Neural Information Processing Systems*, 2021.

[17] Radu Alexandru Dragomir, Mathieu Even, and Hadrien Hendrikx. Fast stochastic bregman gradient methods: Sharp analysis and variance reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2815–2825. PMLR, 18–24 Jul 2021.

[18] Simon Shaolei Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=pW2Q2xLwIMD.

[19] John Duchi and Ryan Rogers. Lower bounds for locally private estimation via communication complexity. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1161–1191, Phoenix, USA, June 2019. PMLR.

[20] John C. Duchi and Martin J. Wainwright. Distance-based and continuum fano inequalities with applications to statistical estimation. 2013.

[21] Mathieu Even and Laurent Massoulié. Concentration of Non-Isotropic Random Tensors with Applications to Learning and Empirical Risk Minimization. *arXiv:2102.04259 [cs, math, stat]*, February 2021. URL http://arxiv.org/abs/2102.04259. arXiv: 2102.04259.

[22] Mathieu Even, Raphaël Berthier, Francis Bach, Nicolas Flammarion, Hadrien Hendrikx, Pierre Gaillard, Laurent Massoulié, and Adrien Taylor. Continuized accelerations of deterministic and stochastic gradient descents, and of gossip algorithms. In *Advances in Neural Information Processing Systems*, 2021.

[23] Mathieu Even, Laurent Massoulié, and Kevin Scaman. Sample optimality and all-for-all strategies in personalized federated and collaborative learning, 2022. URL https://arxiv.org/abs/2201.13097.

[24] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized Federated Learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020.

[25] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19586–19597. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/e32cc80bf07915058ce90722ee17bb71-Paper.pdf.

[26] Felix Grimberg, Mary-Anne Hartley, Sai P. Karimireddy, and Martin Jaggi. Optimal Model Averaging: Towards Personalized Collaborative Learning. *arXiv:2110.12946 [cs, stat]*, October 2021. arXiv: 2110.12946.

[27] Filip Hanzely, Slavomír Hanzely, Samuel Horváth, and Peter Richtarik. Lower Bounds and Optimal Algorithms for Personalized Federated Learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 2304–2315. Curran Associates, Inc., 2020.

[28] Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(none):1 – 6, 2012. doi: 10.1214/ECP.v17-2079. URL https://doi.org/10.1214/ECP.v17-2079.

[29] Yihan Jiang, Jakub Konečný, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning, 2019. URL https://arxiv.org/abs/1909.12488.

[30] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and Open Problems in Federated Learning. *arXiv:1912.04977 [cs, stat]*, December 2019. arXiv: 1912.04977.

[31] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5132–5143. PMLR, November 2020. ISSN: 2640-3498.

[32] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtarik. Tighter theory for local sgd on identical and heterogeneous data. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4519–4529. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/bayoumi20a.html.

[33] Mikhail Khodak, Maria-Florina F Balcan, and Ameet S Talwalkar. Adaptive Gradient-Based Meta-Learning Methods. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[34] Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110 – 133, 2017. doi: 10.3150/15-BEJ730. URL https://doi.org/10.3150/15-BEJ730.

[35] Jakub Konečný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. *arXiv:1610.02527 [cs]*, October 2016. arXiv: 1610.02527.

[36] Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant. Survey of Personalization Techniques for Federated Learning. *arXiv:2003.08673 [cs, stat]*, March 2020. arXiv: 2003.08673.

[37] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020.

[38] Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three Approaches for Personalization with Applications to Federated Learning. *arXiv:2002.10619 [cs, stat]*, July 2020. arXiv: 2002.10619.

[39] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017.

[40] Volodymyr Melnykov and Ranjan Maitra. Finite mixture models and model-based clustering. *Statistics Surveys*, 4(none):80 – 116, 2010. doi: 10.1214/09-SS053. URL https://doi.org/10.1214/09-SS053.

[41] Aritra Mitra, Rayana Jaafar, George J. Pappas, and Hamed Hassani. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 14606–14619. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/7a6bda9ad6ffdac035c752743b7e9d0e-Paper.pdf.

[42] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic Federated Learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4615–4625. PMLR, May 2019. ISSN: 2640-3498.

[43] Angelia Nedich, Alex Olshevsky, and Michael G. Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, May 2018.

[44] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal convergence rates for convex distributed optimization in networks. *Journal of Machine Learning Research*, 20(159):1–31, 2019.

[45] Nathan Srebro, Gregory Shakhnarovich, and Sam Roweis. An investigation of computational and informational limits in gaussian mixture clustering. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 865–872, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143953. URL https://doi.org/10.1145/1143844.1143953.

[46] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50):1517–1561, 2010. URL http://jmlr.org/papers/v11/sriperumbudur10a.html.

[47] Canh T. Dinh, Nguyen Tran, and Josh Nguyen. Personalized Federated Learning with Moreau Envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.

[48] Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7852–7862. Curran Associates, Inc., 2020.

[49] Paul Vanhaesebrouck, Aurélien Bellet, and Marc Tommasi. Decentralized Collaborative Learning of Personalized Models over Networks. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 509–517. PMLR, 20–22 Apr 2017.

[50] Kangkang Wang, Rajiv Mathews, Chloé Kiddon, Hubert Eichner, Françoise Beaufays, and Daniel Ramage. Federated Evaluation of On-device Personalization. *arXiv:1910.10252 [cs, stat]*, October 2019. arXiv: 1910.10252.

[51] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620 – 2648, 2019. doi: 10.3150/18-BEJ1065. URL https://doi.org/10.3150/18-BEJ1065.

[52] Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local sgd for heterogeneous distributed learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6281–6292. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/45713f6ff2041d3fdfae927b82488db8-Paper.pdf.

[53] Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. Salvaging Federated Learning by Local Adaptation. *arXiv:2002.04758 [cs, stat]*, October 2021. arXiv: 2002.04758.

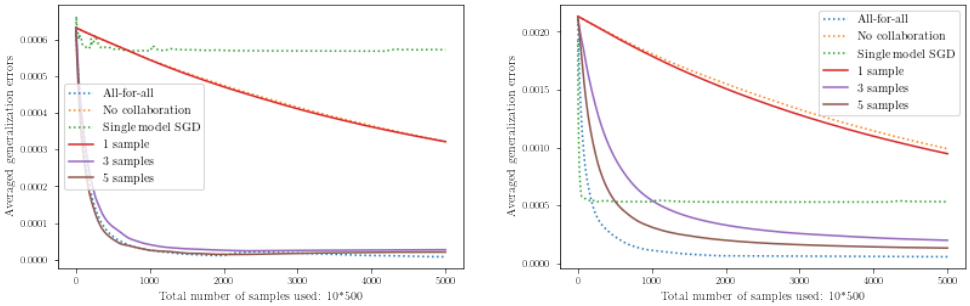
[54] Yuchen Zhang, John Duchi, Michael I Jordan, and Martin J Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] This work is mainly theoretical, and we do not believe it to have any negative societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] , except when proving similar results, we do not repeat the full proof but only sketch the main differences.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] The experiments we ran were consistent through the runs, and are illustrations of our theory.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Numerical illustration of our theory



(a) Agents in unknown clusters

(b) Agents *i.i.d.* distributed

Figure 1: **All-for-all algorithm in practice**

To test the robustness of our theory, we build toy problems from synthetic datasets, placing ourselves in the scenario we considered throughout the paper: a large number of agents with heterogeneous data, that each have too few samples available from their local data distribution in order to reach a small generalization error on their own.

In Figure 1, we consider $N = 500$ agents, and a quadratic loss function $\ell(x, \xi = (a, b)) = \frac{1}{2}(a^\top x - b)^2$, for $x, a \in \mathbb{R}^d$ ($d = 100$) and $\mu \in \mathbb{R}^d$. For $i = 1, \dots, 500$, the distribution $\mathcal{D}_i$ of $\xi_i = (a_i a_i^\top, a_i b_i)$ as a centered Gaussian random variable of covariance matrix Σ_i for a_i, and b_i is the sign of $a_i^\top u$ for some fixed $u \in \mathbb{R}^d$, flipped with probability 0.2. In both figures, each agents have 10 samples available for the optimization phase ($K = 10$ oracle calls), corresponding to a total number of samples used of $N \times K = 5000$. We computed and showed the 500 steps of all ten oracle calls, each step corresponding to the use of the stochastic gradient of a single agent.

The dotted lines represent our baselines. The blue one is the ALL-FOR-ALL algorithm with matrix W exactly as in Corollary 1 with $b_{ij} = \|\Sigma_i - \Sigma_j\|$. The orange dotted line consists in the no-collaboration baseline: each agents performs SGD on its own without sharing information (corresponds to $W = I_N$). The green dotted line corresponds to the "single-model" approach without personalization: one model is trained for all agents, using SGD and all samples from all agents (corresponding to $W = \frac{1}{n}\mathbb{1}\mathbb{1}^\top$). The choice of the algorithm for the single model approach without collaboration is in fact unimportant, since all algorithms would reach the same asymptotic bias here. The full lines (red, violet and brown) correspond to estimating the pairwise distances from empirical distributions (as in Section 6), using respectively $S = 1$, $S = 3$ and $S = 5$ samples.

In Figure 1(a), we consider $M = 10$ (unkonwn) clusters $\mathcal{C}_1, \dots, \mathcal{C}_M$. All $i \in \mathcal{C}_m$ have the same covariance matrix Σ_m, equal to $I_d/\sqrt{d} + e_m e_m^\top$, where e_m is the m-th element of the canonical basis of $\mathbb{R}^d$. In Figure 1(b), $\Sigma_i = \mathrm{Diag}(u_1^{(i)}, \dots, u_1^{(d)})/\sqrt{d}$ where the $(u_\ell^{(i)})$ are *i.i.d.* uniformly distributed in $[0, 1]$. Performing rough estimations of the pairwise distance between agents' local distributions thus appears to be quite robust in both our settings. In the "cluster" setting, this was predicted by our theory, and the numerical results are compelling. In the "*i.i.d.*" setting, using very few samples for the estimation also appears to be very efficient.

B Proof of our lower-bound (Theorem 1)

B.1 General framework to prove lower bounds [1]

The idea is that, when optimizing a function $f(x) = \mathbb{E}\left[\ell(x, \xi)\right]$ and finding a good approximation of a minimizer $x^\star$, we learn some information on the distribution $\mathcal{D}$ over which samples are drawn. In order to prove lower bounds, we construct a loss function ℓ, and distributions $\mathcal{D}_1^\alpha, \dots, \mathcal{D}_N^\alpha$, where α is a random parameter. We argue that minimizing the objective function up to a certain precision gives a good estimator (quantified) of the random seeds α. Then, using Fano inequality, we bound

the efficiency of such an estimator in terms of number of oracle calls, obtaining a lower bound on the sample complexity. This approach is inspired by Agarwal et al. [1], who prove IT-lower bounds for stochastic gradient descent. We adapt their proof technique to the personalized and multi-agent setting.

Constructing difficult loss functions For any two functions $f, g : \mathbb{R}^d \to \mathbb{R}$, we define the discrepancy measure $\rho(f, g)$ as:

$$\rho(f,g) = \inf_{x\in\mathbb{R}^d} \left\{ f(x) + g(x) - \inf_{y\in\mathbb{R}^d} f(y) - \inf_{y\in\mathbb{R}^d} g(y) \right\},$$

which is a pseudo metrics. Now, for a finite set $\mathcal{V}$ of parameters, let $\mathcal{G}(\delta) = \left\{ g_\alpha^\delta \,,\, \alpha \in \mathcal{V} \right\}$ be a set of functions indexed by $\mathcal{V}$, that depend on δ (fixed in the set). The dependency in δ of each $g_\alpha \in \mathcal{G}(\delta)$ is left implicit in the following subsections. We define:

$$\psi(\delta) = \inf_{f,g\in\mathcal{G}(\delta), f\neq g} \rho(f,g)\,.$$

Minimizing is Bernoulli parameters identification The two following lemmas justify that optimizing a function $g_\alpha \in \mathcal{G}(\delta)$ to a precision of order $\psi(\delta)$ is more difficult than estimating the parameter α.

Lemma 1 (Agarwal et al. [1])**.** *For any $x \in \mathbb{R}^d$, there can be at most one function g_α in $\mathcal{G}(\delta)$ such that:*

$$g_\alpha(x) - \inf_{\mathbb{R}^d} g_\alpha < \frac{\psi(\delta)}{3}\,.$$

Lemma 2 (Agarwal et al. [1])**.** *Assume that for some fixed but unknown $\alpha \in \mathcal{V}$ there exists a method $\mathcal{M}_K$ based on the data $\phi = \{X_1, ..., X_K\}$ that returns x^K (function of ϕ) satisfying an error of:*

$$\mathbb{E}\left[g_\alpha(x^K) - \min_{x\in\mathbb{R}^d} g_\alpha(x) \right] < \frac{\psi(\delta)}{9}\,,$$

where the mean is taken over the randomness of both the oracle Φ, the method $\mathcal{M}_K$ and $\alpha \in \mathcal{V}$ if random. Then, there exists a hypothesis test $\hat{\alpha} : \phi \to \mathcal{V}$ such that:

$$\max_{\alpha\in\mathcal{V}} \mathbb{P}_\phi\big(\hat{\alpha} \neq \alpha\big) \leqslant \frac{1}{3}\,.$$

Suppose now that the parameter α in the previous Lemma is chosen uniformly at random in $\mathcal{V}$. Let $\hat{\alpha} : \phi \to \mathcal{V}$ be a hypothesis test estimating α. By Fano inequality [12], we have:

$$\mathbb{P}\big(\hat{\alpha} \neq \alpha\big) \geqslant 1 - \frac{I\big(\phi, \alpha\big) + \ln(2)}{\ln(|\mathcal{V}|)}\,, \tag{4}$$

where $I\big(\phi, \alpha\big)$ is the mutual information between ϕ and α, that we need to upper-bound. Combining Fano inequality with Lemmas 1 and 2, fixing a target error $\varepsilon = \psi(\delta)$, we obtain a lower bound on the K_ε the number of oracle calls required to reach an ε generalization error:

$$\frac{1}{3} \geqslant \mathbb{P}_\phi\big(\hat{\alpha} \neq \alpha\big) \geqslant 1 - \frac{I\big(\phi_{K_\varepsilon}, \alpha\big) + \ln(2)}{\ln(|\mathcal{V}|)}\,,$$

where ϕ_{K_ε} is the information contained in K_ε oracle calls. If we have an equality of the form $I\big(\phi_{K_\varepsilon}, \alpha\big) = K_\varepsilon I\big(\phi_1, \alpha\big)$, this gives:

$$K_\varepsilon \geqslant \frac{\frac{2}{3}\ln(|\mathcal{V}|) - \ln(2)}{I(\phi_1, \alpha)}\,. \tag{5}$$

Playing with the different parameters $\delta, \alpha, \mathcal{V}$ gives lower bounds. We refer the interested reader to Chapter 2 in Cover and Thomas [12] for Fano inequality and mutual information.

B.2 Applying this to prove Theorem 1

For simplicity, assume that $r^2 = d$ and $\sigma^2 = 1$. Let $\delta > 0$ a free parameter. Let $\mathcal{V} = \{\alpha^1, \dots, \alpha^L\} \subset \{-1, 1\}^d$ be a subset of the hypercube such that for all $k \neq l$,

$$\frac{1}{2}\sum_{i=1}^{d} |\alpha_i^k - \alpha_i^l| \geqslant \frac{d}{4}\,,$$

i.e. $\mathcal{V}$ is a $d/4$-packing of the hypercube. We know that we can set $|\mathcal{V}| \geqslant (2/\sqrt{e})^{d/2}$. Without loss of generality, we prove a lower bound in the case where the agent that desires to minimize its local function is indexed by 1.

Let:

$$\ell(x, \xi) = \frac{1}{2}\|x - \xi\|^2\,,$$

for $x, \xi \in \mathbb{R}^d$ and, for fixed $\delta > 0$ and any $\alpha \in \mathcal{V}$:

$$g_\alpha(x) = \frac{1}{2d}\sum_{k=1}^{d}\Big(x_k^2 + 1 - 2\big(\frac{1}{2} + \alpha_k\delta\big)x_k\Big) \quad , x \in \mathcal{X}\,.$$

We keep the same notations as last subsection $(\psi(\delta), \rho)$. We have:

$$\rho(g_\alpha, g_\beta) = \frac{\delta^2}{d}\sum_{k=1}^{d}|\alpha_k - \beta_k|\,,$$

leading to $\psi(\delta) \geqslant \delta^2/4$ since $\mathcal{V}$ is a $d/4$-packing of the hypercube.

For any $i = 1, \dots, N$, let $\mathcal{D}_i$ be the probability distribution on $\{0, 1\}^d$ of the following random variable:

$$\mathrm{Ber}\big(\frac{1}{2} + \delta_i\alpha_k\big)\epsilon_k \quad \text{where} \quad \delta_i = (\delta - b_{i1})^+\,,$$

where $s^+ = \max(0, s)$ for $s \in \mathbb{R}$, k is taken uniformly at random in $\{1, \dots, d\}$, (ϵ_k) is the canonical basis of $\mathbb{R}^d$, and $Ber(p)$ is a Bernoulli random variable, independent of k.

The mutual information is thus, in our case:

$$I(\phi_K, \alpha) \leqslant C_1 K \mathcal{N}_1^\delta(\sqrt{b})\delta^2\,,$$

where we use the fact that $I\left(\mathrm{Ber}\big(\frac{1}{2} + \mathbb{1}_{b_{1i}\leqslant\delta}\alpha_k\delta_i\big), \alpha_k\right) \leqslant C_1\delta^2$ for some constant $C_1 > 0$, for $\delta_i \leqslant 1/4$. Setting the target precision as $\varepsilon = \delta^2/4$, we obtain:

$$K_\varepsilon \geqslant C'\frac{d}{\varepsilon\mathcal{N}_1^\varepsilon(4b)}\,.$$

The loss function and distributions built verify our regularity assumptions for $\mu = 1/d$, $L = 1/d$, noise $\sigma^2 \leqslant 1$.

We first verify that for all $1 \leqslant j, k \leqslant N$, we have $f_j(x_k^\star) - f_j(x_j^\star) \leqslant b_{kj}^2$. We first notice that $x_j^\star = \frac{1}{d}\big(\frac{1}{2} + \delta_j\alpha_l\big)_{1\leqslant l\leqslant d}$, so that:

$$\begin{aligned} f_j(x_k^\star) - f_j(x_j^\star) &= \frac{1}{d}\big\|x_j^\star - x_k^\star\big\|^2 \\ &= (\delta_i - \delta_k)^2 \\ &\leqslant (b_{1j} - b_{1k})^2 \\ &\leqslant |b_{1j} - b_{1k}|^2 \\ &\leqslant b_{jk}^2\,, \end{aligned}$$

since the weights b verify the triangle inequality. Under the assumptions of Theorem 1 on $\mathcal{H}$, we have, in terms of distribution-based distances:

$$d_{\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_j) \leqslant |\delta_i - \delta_j| \leqslant b_{ij}\,.$$

The minimum of each g_α is attained at $x^\alpha = \frac{1}{2} + \delta\alpha$, we thus need to assume that r is of order $\sqrt{d}$, and a rescaling leads to the dependency in r. The dependency in σ^2 for $\sigma^2 > 1$ is obtained by taking $\mathcal{D}_i' = \mathrm{Ber}(1/\sigma^2)\sigma^2\mathcal{D}_i$. In this case, we have a noise amplitude of order σ^2 instead of order 1, and a factor $1/\sigma^2$ appears in the mutual information.

C Proof of Theorem 2

Proof of Theorem 2. We begin by proving the following descent lemma.

Lemma 3. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and μ-strongly convex and $G : \mathbb{R}^d \to \mathbb{R}$ be differentiable. Consider the iterates generated by:*

$$y^{k+1} = y^k - \eta g^k \,,$$

where $\mathbb{E}\left[g_k | y^k\right] = \nabla G(y^k)$ and $\mathbb{E}\left[\left\|g^k - \nabla G(y^k)\right\|^2 | y^k\right] \leqslant \sigma_g^2$. Then, we have, where $y^\star$ minimizes F, as long as $\eta \leqslant 1/L$:

$$\mathbb{E}\left[F(y^{k+1}) - F(y^\star)\right] \leqslant (1-\eta\mu)\mathbb{E}\left[F(y^k) - F(y^\star)\right] + \frac{\eta}{2}\mathbb{E}\left[\left\|\nabla F(y^k) - \nabla G(y^k)\right\|^2\right] + \frac{\eta^2\sigma_g^2 L}{2}$$

Proof Lemma 3. We use smoothness of F:

$$\mathbb{E}\left[F(y^{k+1}) - F(y^k)\right] \leqslant -\eta\mathbb{E}\left[\langle g^k, \nabla F(y^k)\rangle\right] + \frac{\eta^2 L}{2}\mathbb{E}\left[\left\|g^k\right\|^2\right] \,.$$

Then, using $\mathbb{E}\left[\left\|g^k\right\|^2\right] \leqslant \mathbb{E}\left[\left\|\nabla G(y^k)\right\|^2\right] + \sigma_g^2$, and $-\eta\mathbb{E}\left[\langle g^k, \nabla F(y^k)\rangle\right] = -\eta\mathbb{E}\left[\langle \nabla G(y^k), \nabla F(y^k)\rangle\right] = -\frac{\eta}{2}\mathbb{E}\left[\left\|\nabla G(y^k)\right\|^2 + \left\|\nabla F(y^k)^2\right\| - \left\|\nabla G(y^k) - \nabla F(y^k)\right\|^2\right]$, we obtain that:

$$\begin{aligned}\mathbb{E}\left[F(y^{k+1}) - F(y^k)\right] \leqslant -\frac{\eta}{2}\mathbb{E}\left[\left\|\nabla F(y^k)\right\|^2\right] - \frac{\eta}{2}\mathbb{E}\left[\left\|\nabla G(y^k)\right\|^2\right](1-\eta L) \\ + \frac{\eta}{2}\mathbb{E}\left[\left\|\nabla F(y^k) - \nabla G(y^k)\right\|^2\right] + \frac{\sigma_g^2\eta^2 L}{2} \,.\end{aligned}$$

Finally, we conclude using $-\frac{\eta}{2}\mathbb{E}\left[\left\|\nabla F(y^k)\right\|^2\right] \leqslant \eta\mathbb{E}\left[F(y^k) - F(y^\star)\right]$ and $\eta < 1/L$. □

Let $i \in [N]$. To prove Theorem 2, we now use Lemma 3 to study the sequence $y^k = x_i^k$ with $F = f_i$, $G = \sum_j \lambda_{ij} f_j := f^\lambda$ and $g^k = \sum_j \lambda_{ij} g_j^k(x_i^k)$. For all $x \in \mathbb{R}^d$, using Assumption 3:

$$\begin{aligned}\left\|\nabla f_i(x) - \nabla f^\lambda(x)\right\|^2 &\leqslant \sum_{j=1}^N \lambda_{ij}\|f_i(x) - f_j(x)\|^2 \\ &= \sum_{j=1}^N \lambda_{ij}\|\mathbb{E}\left[\nabla_x \ell(x,\xi_i)\right] - \mathbb{E}\left[\nabla_x \ell(x,\xi_j)\right]\|^2 \\ &\leqslant \sum_{j=1}^N \lambda_{ij} b_{ij}^2 \,,\end{aligned}$$

since $\nabla_x \ell(x,\cdot) \in \mathcal{H}$ and $d_{\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_j) \leqslant b_{ij}$. Then, using the independence (conditionally on x_i^k) of the $(g_j^k)_j$ and $\mathbb{E}\left[g_j^k(x_i^k) | x_i^k\right] = \nabla f_j(x_i^k)$:

$$\mathbb{E}\left[\left\|\sum_j \lambda_{ij} g_j^k(x_i^k) - \sum_j \lambda_{ij}\nabla f_j(x_j^k)\right\|^2\right] = \sum_j \mathbb{E}\left[\left\|\lambda_{ij}\left(g_j^k - \nabla f_j(x_j^k)\right)\right\|^2\right] \leqslant \sum_j \lambda_{ij}^2\sigma^2 \,.$$

Consequently,

$$\mathbb{E}\left[f_i(x_i^{k+1}) - f_i(x_i^\star)\right] \leqslant (1-\eta\mu)\mathbb{E}\left[f_i(x_i^k) - f_i(x_i^\star)\right] + \frac{\eta}{2}\sum_{j=1}^N \lambda_{ij} b_{ij}^2 + \frac{\eta^2\sigma^2 L}{2}\sum_{j=1}^N \lambda_{ij}^2 \,.$$

Writing $H^k = (1-\eta\mu)^{-k}\mathbb{E}\left[f_i(x_i^k) - f_i(x_i^\star)\right]$, unrolling the recursion leads to:

$$H^K \leqslant H^0 + \frac{\eta}{2}\sum_{k<K}(1-\eta\mu)^{-k}\sum_{j=1}^N \lambda_{ij} b_{ij}^2 + \sum_{k<K}(1-\eta\mu)^{-k}\frac{\eta^2\sigma^2 L}{2}\sum_{j=1}^N \lambda_{ij}^2 \,.$$

Finally, using $\sum_{k<K}(1-\eta\mu)^{-k} \leqslant \frac{(1-\eta\mu)^{-K}}{\eta\mu}$, we have:

$$\mathbb{E}\left[f(x_i^K) - f(x_i^\star)\right] \leqslant (1-\eta\mu)^K (f(x_i^0) - f(x_i^\star)) + \frac{\eta\sigma^2 L}{2\mu}\sum_{j=1}^N \lambda_{ij}^2 + \frac{1}{2\mu}\sum_{j=1}^N \lambda_{ij} b_{ij}^2 \,.$$

Using $(1-\eta\mu)^K \leqslant e^{-K\eta\mu}$, we optimize of η. For

$$\eta_i = \min\left\{\frac{1}{2L}\,,\, \frac{1}{\mu K}\ln\left(\frac{2\mu^2 K(f(x_i^0) - f(x_i^\star))}{\sigma^2 L \sum_j \lambda_{ij}^2}\right)\right\}, \tag{6}$$

we obtain:

$$\begin{aligned}\mathbb{E}\left[f(x_i^K) - f(x_i^\star)\right] &\leqslant (f(x_i^0) - f(x_i^\star))e^{-K/\kappa} \\ &\quad + \frac{\sigma^2 L}{\mu^2 K}\ln\left(\frac{2\mu^2 K(f(x_i^0) - f(x_i^\star))}{\sigma^2 L \sum_j \lambda_{ij}^2}\right)\sum_j \lambda_{ij}^2 + \frac{1}{2\mu}\sum_{j=1}^N \lambda_{ij} b_{ij}^2 \,,\end{aligned}$$

leading to the first part of Theorem 2. For the second part, we simply plug the expression of λ_{ij} in the proven formula. □

D Proof of Theorem 3

We recall that for a stochastic matrix Λ, we defined

$$f^\Lambda(y) = \frac{1}{N}\sum_{i=1}^N f_i\big(\sum_{j=1}^N \lambda_{ij} y_j\big)\,, \quad y = (y_1, \dots, y_N) \in \mathbb{R}^{N\times d}\,.$$

Then, y^Λ is defined as a minimizer of f^Λ, and we write $x^\star = (x_1^\star, \dots, x_N^\star)$ where $x_i^\star$ is the minimizer of f_i.

We first begin with the following simple lemmas.

Lemma 4. *If Assumption 3 (weak version) holds, then for all $i, j = 1, \dots, N$, we have:*

$$f_i(x_j^\star) - f_i(x_i^\star) \leqslant \frac{b_{ij}^2}{2\mu}\,.$$

Proof. Using strong-convexity of f_i and $\nabla f_i(x_i^\star)$:

$$\begin{aligned}f_i(x_j^\star) - f_i(x_i^\star) &\leqslant \frac{1}{2\mu}\left\|\nabla f_i(x_j^\star)\right\|^2 \\ &= \frac{1}{2\mu}\left\|\nabla f_i(x_j^\star) - \nabla f_i(x_i^\star)\right\|^2 \\ &= \frac{1}{2\mu}\left\|\mathbb{E}\left[\nabla_x \ell(x_j^\star, \xi_i)\right] - \mathbb{E}\left[\nabla_x \ell(x_j^\star, \xi_i)\right]\right\|^2 \\ &\leqslant \frac{b_{ij}^2}{2\mu}\,,\end{aligned}$$

where the last inequality is deduced using the weak version of Assumption 3. □

Lemma 5. *If Λ is a stochastic matrix,*

$$f^\Lambda(y^\Lambda) - \bar{f}(x^\star) \leqslant \frac{1}{N}\sum_{1\leqslant i,j\leqslant N} \lambda_{ij}\big(f_i(x_j^\star) - f_i(x_i^\star)\big)\,.$$

Proof. Writing the optimality of y^Λ gives:

$$\begin{aligned}f^\Lambda(y^\Lambda) &\leqslant f^\Lambda(x^\star) \\ &= \frac{1}{N}\sum_i f_i(\sum_j \lambda_{ij} x_j^\star) \\ &\leqslant \frac{1}{N}\sum_{1\leqslant i,j\leqslant N} \lambda_{ij} f_i(x_j^\star)\,,\end{aligned}$$

where we used convexity of each f_i. Then, subtracting $\bar{f}(x^\star)$ and using stochasticity of Λ:

$$f^\Lambda(y^\Lambda) - \bar{f}(x^\star) \leqslant \frac{1}{N} \sum_{1 \leqslant i,j \leqslant N} \lambda_{ij}(f_i(x_j^\star) - f_i(x_i^\star)) \,.$$

□

We are now armed to prove Theorem 3.

Proof. We have the following bias-variance decomposition, where the inequality is a consequence of Lemma 5:

$$\begin{aligned} F^k &= \bar{f}(y^k) - \bar{f}(x^\star) \\ &= f^\Lambda(y^k) - f^\Lambda(y^\Lambda) + f^\Lambda(y^\Lambda) - \bar{f}(x^\star) \\ &\leqslant f^\Lambda(y^k) - f^\Lambda(y^\Lambda) + \frac{1}{N} \sum_{1 \leqslant i,j \leqslant N} \lambda_{ij} \frac{b_{ij}^2}{2\mu} \,. \end{aligned}$$

We thus need to upper-bound the optimization term $f^\Lambda(y^k) - f^\Lambda(y^\Lambda)$. We recall that y^k verifies the recursion:

$$y^{k+1} = y^k - \eta \nabla G_\Lambda^k(y^k) \,,$$

for

$$G_\Lambda^k(y) = \frac{1}{N} \big(\sum_{i=1}^N \lambda_{ij} g_i^k((\Lambda y^k)_i) \big)_{1 \leqslant j \leqslant N} \,,$$

that verifies:

$$\begin{aligned} &\mathbb{E}\left[G_\Lambda^k(y)\right] = \nabla f^\Lambda(y) \,, \\ &\mathbb{E}\left[\left\|G_\Lambda^k(y) - \nabla f^\lambda(y)\right\|^2\right] \leqslant \frac{\sigma^2}{N^2} \sum_{1 \leqslant i,j \leqslant N} \lambda_{ij}^2 \,. \end{aligned}$$

The function f^Λ is however not necessarily strongly convex. However, since $\nabla^2 f^\Lambda(y) = \Lambda^\top \nabla^2 \bar{f}(\Lambda y) \Lambda$ and $\bar{f}$ is L/N-smooth and μ/N-strongly convex, f^Λ is L/N-relatively smooth and μ/N-relatively strongly convex [2] with respect to $\frac{1}{2}\|y\|_W^2 = \frac{1}{2} y^\top W y$. Note also that the spectral radius of W is 1, since Λ is stochastic. Instead of using stochastic Bregman gradient descent (*e.g.* Dragomir et al. [17]), we use Lemma 6 that we prove at the end of the paper: classical SGD that naturally generalizes to relative smoothness and strong convexity assumptions, when the mirror map is quadratic. This leads to:

$$\mathbb{E}\left[f^\Lambda(y^k) - f^\Lambda(y^\Lambda)\right] \leqslant (f^\Lambda(y^0) - f^\Lambda(y^\Lambda)) e^{-\frac{K}{2\kappa}} + \frac{\kappa \sigma^2}{K \mu N} \ln\left(\frac{2\mu^2 K (f^\Lambda(y^0) - f^\Lambda(y^\Lambda))}{\sigma^2 L \sum_{i,j} \frac{1}{N} \lambda_{ij}^2} \right) \sum_{1 \leqslant i,j \leqslant N} \lambda_{ij}^2 \,,$$

for a choice of stepsizes of:

$$\eta = \min\left\{ \frac{1}{2L} \,, \frac{1}{\mu K} \ln\left(\frac{2\mu^2 K (f^\Lambda(y^0) - f^\Lambda(y^\Lambda))}{\sigma^2 L \sum_{i,j} \frac{1}{N} \lambda_{ij}^2} \right) \right\} , \tag{7}$$

concluding the proof. □

E Proof of Theorem 4

We first start by recalling that, for a Σ-subgaussian random variable ξ, using Theorem 1 from [28], we have for any $t \geqslant 0$:

$$\mathbb{P}\left(\|\xi\|^2 \geqslant d_{\text{eff}} \nu^2 + 2\sqrt{\|\Sigma\|_2 t} + 2\nu^2 t \right) \leqslant e^{-t} \,.$$

Consequently, for $u \geqslant \max(\nu^2 d_{\text{eff}}, \|\Sigma\|_2^2/\nu^2)$, we have:

$$\mathbb{P}\Big(\|\xi\|^2 \geqslant 4u\Big) \leqslant e^{-\frac{u}{2\nu^2}} .$$

Remarking that $\frac{\|\Sigma\|_2^2}{\nu^2} = \nu^2 \sum_k \frac{\nu_k^4}{\nu^4} \leqslant \nu^2 \sum_k \frac{\nu_k^2}{\nu^2} = \nu^2 d_{\text{eff}}$, where ν_k^2 are the eigenvalues of Σ, this condition on u is in fact $u \geqslant \nu^2 d_{\text{eff}}$.

Proof of our Theorem. From Theorem 3, we have, conditionally on the S samples used in estimating biases:

$$\mathbb{E}\left[F^K|\hat{\Lambda}\right] \leqslant F^0 e^{-\frac{K}{2\kappa}} + \tilde{\mathcal{O}}\left(\frac{\kappa\sigma^2}{K\mu N}\sum_{1\leqslant i,j\leqslant N}\hat{\lambda}_{ij}^2\right) + \frac{1}{N}\sum_{1\leqslant i,j\leqslant N}\hat{\lambda}_{ij}\frac{b_{ij}^2}{2\mu} .$$

We hence need to bound $\mathbb{E}\left[\sum_{i,j}\hat{\lambda}_{ij}^2\right]$ and $\mathbb{E}\left[\sum_{i,j}\hat{\lambda}_{ij}b_{ij}^2\right]$, and start by assuming that $u \geqslant \frac{4}{S}\nu^2 d_{\text{eff}}$. First, denoting $b_{\max} = \max_{i,j} b_{ij}$, we have:

$$\begin{aligned}
\mathbb{E}\left[\sum_{i,j}\hat{\lambda}_{ij}b_{ij}^2\right] &= \sum_{i,j}\mathbb{E}\left[\hat{\lambda}_{ij}b_{ij}^2\right] \\
&= \sum_{i,j:\, b_{ij}^2>4u}\mathbb{E}\left[\hat{\lambda}_{ij}b_{ij}^2\right] + \sum_{i,j:\, b_{ij}^2\leqslant 4u}\mathbb{E}\left[\hat{\lambda}_{ij}b_{ij}^2\right] \\
&\leqslant b_{\max}^2\mathbb{P}\Big(\hat{b}_{ij}^2 \leqslant u|b_{ij}^2 > 4u\Big) + 4u^2\mathbb{1}_{\{4u^2\geqslant\Delta\}} .
\end{aligned}$$

Using a triangle inequality, that gives us $2\|\hat{\mu}_i - \hat{\mu}_j - \mathbb{E}\left[\hat{\mu}_i - \hat{\mu}_j\right]\|^2 \geqslant b_{ij}^2 - 2\hat{b}_{ij}^2$, $\mathbb{P}\Big(\hat{b}_{ij}^2 \leqslant u|b_{ij}^2 > 4u\Big) \leqslant \mathbb{P}\Big(\|\hat{\mu}_i - \hat{\mu}_j - \mathbb{E}\left[\hat{\mu}_i - \hat{\mu}_j\right]\|^2 \geqslant \frac{b_{ij}^2-2u}{2}\Big)$. Then, since each ξ_i and ξ_j are Σ-subgaussian (and independent), $\hat{\mu}_i - \hat{\mu}_j - \mathbb{E}\left[\hat{\mu}_i - \hat{\mu}_j\right]$ is $4\Sigma/S$ subgaussian[2], so that using our assumption on u, we can use Theorem 1 of Hsu et al. [28]:

$$\mathbb{P}\Big(\hat{b}_{ij}^2 \leqslant u|b_{ij}^2 > 4u\Big) = \mathbb{P}\Big(\hat{b}_{ij}^2 \leqslant u|b_{ij}^2 > \max(4u,\Delta^2)\Big) \leqslant 2e^{-\frac{S\max(2u,\Delta^2-2u)}{8\nu^2}} .$$

Then, for $1 \leqslant i \leqslant N$, let $\hat{\mathcal{N}}_i = \sum_{j=1}^N \mathbb{1}_{\{\hat{b}_{ij}\leqslant u\}}$. Fix $1 \leqslant m \leqslant M$. We have:

$$\begin{aligned}
\mathbb{P}\Big(\forall i \in \mathcal{C}_m, \|\hat{\mu}_i - \mu_m\|^2 \leqslant u\Big) &= 1 - \mathbb{P}\Big(\exists i \in \mathcal{C}_m, \|\hat{\mu}_i - \mu_m\|^2 > u\Big) \\
&\geqslant 1 - 2Ce^{-\frac{Su}{8\nu^2}} ,
\end{aligned}$$

so that $\mathbb{E}\left[\sum_{i\in\mathcal{C}_M}\sum_{1\leqslant j\leqslant N}\lambda_{ij}^2\right] = \mathbb{E}\left[\frac{1}{C}\sum_{i\in\mathcal{C}_m}\frac{1}{\hat{\mathcal{N}}_i}\right] \leqslant \frac{1}{C} + 2Ce^{-\frac{Su}{8\nu^2}}$, concluding the proof. We then prove the resulting corollary by taking $u = \Delta/4$, and the condition on u translates into $S \geqslant 16\frac{\nu^2 d_{\text{eff}}}{\Delta^2}$.

□

F SGD under strong-convexity and smoothness assumptions

We recall the following well-known result.

[2] indeed, using the subgaussian norm $\|\cdot\|_{\psi_2}$, for real-valued independent random variables $X_1, \ldots, X_S$ and any $\beta > 0$, $\mathbb{E}\left[e^{\beta\frac{1}{S}\sum_{s=1}^S X_s}\right] = \prod_s \mathbb{E}\left[e^{\frac{\beta}{S}X_s}\right] \leqslant \prod_s \mathbb{E}\left[e^{C_{\psi_2}\frac{\beta^2}{S^2}\|X_s\|_{\psi_2}}\right] = \mathbb{E}\left[e^{C_{\psi_2}\frac{\beta^2}{S^2}\sum_{s=1}^S\|X_s\|_{\psi_2}}\right]$, so that $\left\|\frac{1}{S}\sum_{s=1}^S X_s\right\|_{\psi_2} \leqslant \frac{1}{S^2}\sum_{s=1}^S\|X_s\|_{\psi_2}$, and we apply this to the random variables $\hat{\mu}_i\top y$

Lemma 6 (SGD, s.c. and smooth)**.** *Define $\|x\|_A^2 = x^\top A x$ for some non-negative and symmetric matrix A. Let $f : \mathcal{X} \to \mathbb{R}$ μ-relatively strongly convex and L-relatively smooth with respect to $\frac{1}{2}\|x\|_A^2$. Let $(f_t, g_t)_{t\geqslant 0}$ be first order oracle calls such that for all $t \geqslant 0$:*

$$\forall x \in \mathcal{X}\,, \quad \begin{cases} \mathbb{E}\left[f_t(x)\right] = f(x)\,, \\ \mathbb{E}\left[g_t(x)\right] = \nabla f(x)\,, \\ \mathbb{E}\left[\|g_t(x) - \nabla f(x)\|^2\right] \leqslant \sigma^2\,, \end{cases}$$

for some $\sigma > 0$. Let L_A be the largest eigenvalue of A, and assume that $L_A \leqslant 1$ (our result generalizes to any L_A). Let $(x_t)_{t\geqslant 0}$ be generated with:

$$\forall t \geqslant 0\,, \quad x^{t+1} = x^t - \eta g_t(x^t)\,,$$

for a fixed stepsize $\frac{1}{2L} \geqslant \eta > 0$, and assume that all the iterates lie in $\mathcal{X}$. Assume that f is minimized over $\mathcal{X}$ at some interior point $x^\star$. We have for any $T > 0$:

$$\mathbb{E}\left[f(x^T) - f(x^\star)\right] \leqslant e^{-\eta\mu T}\big(f(x^0) - f(x^\star)\big) + \frac{\eta L \sigma^2}{\mu}\,.$$

For fixed $T > 0$, setting $\eta = \min\big(1/(2L), \frac{1}{\mu T}\ln(\frac{f_0\mu^2 T}{L\sigma^2})\big)$ gives:

$$\mathbb{E}\left[f(x^T) - f(x^\star)\right] \leqslant e^{-\frac{\mu}{2L}T}\big(f(x^0) - f(x^\star)\big) + \frac{L\sigma^2}{\mu^2 T}\ln\big(\frac{f_0\mu^2 T}{L\sigma^2}\big)\,.$$

Thus, for fixed target precision $\varepsilon > 0$, using stepsize $\eta_\varepsilon = \min\big(\frac{\mu\varepsilon}{2L\sigma^2}, \frac{1}{2L}\big)$ and setting $T_\varepsilon = \lceil \ln\big(\varepsilon^{-1}(f(x^0) - f(x^\star))\big)\frac{1}{\eta_\varepsilon\mu}\rceil$, we have:

$$f\left(\frac{1}{T_\varepsilon}\sum_{t<T_\varepsilon} x^t\right) - f(x^\star) \leqslant \varepsilon\,,$$

with a number of oracle calls

$$T_\varepsilon \leqslant \max\left(\frac{2L\sigma^2}{\varepsilon\mu^2}, \frac{2L}{\mu}\right)\ln\big(\varepsilon^{-1}(f(x^0) - f(x^\star))\big)\,.$$

Proof. For some $t \geqslant 0$, denoting $f_t = \mathbb{E}\left[f(x^{t+1}) - f(x^\star)\right]$, using relative smoothness, unbiasedness of the stochastic gradients and then relative strong convexity:

$$\begin{aligned} f_{t+1} - f_t &\leqslant -\eta\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + \frac{\eta^2 L}{2}\mathbb{E}\left[\|g_t\|_A^2\right] \\ &\leqslant -\eta\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + \frac{\eta^2 L L_A}{2}\mathbb{E}\left[\|g_t\|^2\right] \\ &\leqslant -\eta\left(1 - \frac{\eta L L_A}{2}\right)\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + \frac{\eta^2 L L_A \sigma^2}{2}\,. \end{aligned}$$

Using relative strong convexity of f, we have:

$$\begin{aligned} \left\|\nabla f(x^t)\right\|^2 &\geqslant \frac{1}{L_A}\left\|\nabla f(x^t)\right\|_A^2 \\ &\leqslant \frac{2\mu}{L_A} f_t\,, \end{aligned}$$

yielding, for $\eta < 1/(LL_A)$

$$f_{t+1} - f_t \leqslant -2\eta\frac{\mu}{L_A} f_t + \frac{\eta^2 L L_A \sigma^2}{2}\,.$$

Then, for some $T > 0$ and since $L_A \leqslant 1$, sum the above inequality multiplied by $(1-\eta\mu)^{-t-1}$:

$$\begin{aligned} \sum_{0\leqslant t\leqslant T-1}(1-\eta\mu)^{-t-1} f_{t+1} - (1-\eta\mu)^{-t} f_t &\leqslant \frac{\eta^2 L\sigma^2}{2}\sum_{0\leqslant t\leqslant T-1}(1-\eta\mu)^{-t-1} \\ &\leqslant \frac{\eta^2 L\sigma^2}{2}\frac{(1-\eta\mu)^{-t-1}}{\eta\mu}\,, \end{aligned}$$

leading to the desired result. □